ReCALL **22**(1): 83–102. 2010 © European Association for Computer Assisted Language Learning
doi:10.1017/S0958344009990218

Utilizing lexical data from a Web-derived corpus to expand productive collocation knowledge

SHAOQUN WU AND IAN H. WITTEN
Computer Science Department, University of Waikato, New Zealand
(email: shaoqun@cs.waikato.ac.nz; ihw@cs.waikato.ac.nz)

MARGARET FRANKEN
School of Education, University of Waikato, New Zealand
(email: franken@waikato.ac.nz)

Abstract

Collocations are of great importance for second language learners, and a learner's knowledge of them plays a key role in producing language fluently (Nation, 2001: 323). In this article we describe and evaluate an innovative system that uses a Web-derived corpus and digital library software to produce a vast concordance and present it in a way that helps students use collocations more effectively in their writing. Instead of live search we use an off-line corpus of short sequences of words, along with their frequencies. They are preprocessed, filtered, and organized into a searchable digital library collection containing 380 million five-word sequences drawn from a vocabulary of 145,000 words. Although the phrases are short, learners can browse more extended contexts because the system automatically locates sample sentences that contain them, either on the Web or in the British National Corpus. Two evaluations were conducted: an expert user tested the system to see if it could generate suitable alternatives for given text fragments, and students used it for a particular exercise. Both suggest that, even within the constraints of a limited study, the system could and did help students improve their writing.

Keywords: concordancing, concordancers, collocation, collocation knowledge, web corpus, data-driven

1 Introduction

Corpus linguistics has moved beyond the realm of pure linguistics and become of interest to those involved in language teaching and learning. As Gabrielatos (2005) states, "'Corpus' has now become one of the new language teaching catchphrases, and both teachers and learners alike are increasingly becoming consumers of corpus-based educational products, such as dictionaries and grammars".

Most corpora are based on particular domains, genres, or collections of certain types of documents from which recurrent phrases and grammatical patterns can easily be retrieved (Stubbs & Barth, 2003). A corpus is therefore a particularly productive

context in which to study collocations, a notoriously challenging aspect of English productive use even for quite advanced learners (Bishop, 2004; Nesselhauf, 2003).

We think of a collocation in the same way as expressed by Benson, Benson and Ilson (1986: ix): "In any language, certain words combine with certain other words or grammatical constructions. These recurrent, semi-fixed combinations, or collocations, can be divided into two groups: grammatical collocations and lexical collocations". We focus on one particular category of collocations, lexical collocations, "which have structures such as the following: verb + noun, adjective + noun, noun + verb. Noun + noun, adverb + adjective, adverb + verb" (Benson *et al.*, 1986: ix). Wei supports this position, arguing that "such an approach to collocation systematically incorporates syntax into a predominantly semantic and lexical construct, thus encompassing a wider range of data" (Wei, 1999: 4).

The article describes and evaluates an innovative system that utilizes a Web-derived corpus and digital library software to produce concordance results aimed at helping students use collocations more effectively. The corpus is enormous: it contains about 145,000 unique words, 14 million two-grams, 420 million three-grams, 500 million four-grams and 380 million five-grams. This is used to form two digital library collections: phrases and collocations, the latter being segmented into different grammatical patterns. Both phrases and collocations are presented to users in order of occurrence frequency, which prioritizes commonly used ones. It is essential to avoid overwhelming users with a superabundance of choice, and the system design focuses on providing suitably targeted searching and browsing facilities that help actual users doing real language tasks. In order to provide a realistic context of use, we recruited language learners from an International English Language Testing System (IELTS) writing preparation class. Each wrote an essay and then used the system to correct it. Results were extremely encouraging.

We begin by exploring the nature of concordancers, their effect on learners' language usage, and the use of the Web as a corpus. We utilize an off-line corpus generated and supplied by Google, and explain how we process and filter this before making two digital library collections. The facilities for searching and browsing are the key to effective use of this system, and we describe the design and implementation of these next. Finally we give a comprehensive account of the evaluation and the results obtained.

2 The nature of concordancers

A concordancer is "a piece of software, either installed on a computer or accessed through a website, which can be used to search, access and analyse language from a corpus" (Peachey, 2005: Section 1). Concordancers are among the most frequently used tools to explore corpora, specifically with a view to examining collocational use. They make it possible for students to obtain, organize, and study real-language data derived from corpora. However, it is important to consider the nature of the concordancer, because it mediates the corpus for the user.

A typical concordancer allows users to specify words or phrases and search for examples of their use by providing a context in which the searched item appears. However, concordancers differ in the way in which data is presented. Not all concordance

results are easily navigated and analysed by learners. For example, learners are easily overwhelmed by the vast number of concordances returned when searching for a common word. Concordancers also differ in the size of the unit of analysis they present. Some present short snippets of text, while others present items in paragraph-length units. Concordancers can also limit the items that are retrievable.

One accessible and user-friendly concordancer available on the Web is the Compleat Lexical Tutor from Université du Québec à Montréal (Cobb, n.d.). Using this website tool, students can enter a word and explore what words are most likely to occur after the core item and/or before it. They specify a keyword to search for, and select which of a number of different corpora to search in. They can also associate another word with the keyword, specifying a position – left, right or any. The search results are chunks of text (constrained by line width) that contain the keyword and, if specified, the associated word.

Collocation dictionaries are another language learning and teaching resource that relies on concordancing. One example is the online Cobuild Concordance and Collocations Sampler, which allows learners to search for collocations of a particular word. However, it only provides a demonstration facility and the returned result is the list of words occurring on either side of the target word, along with some statistical data. The collocations offered are intended for lexicologists, not applied linguists, and may not be particularly useful for learning.

It is clear that concordancers need to present results in a way that is both accessible and relevant to learners. They must provide sufficient and varied language data, and in combinations that are flexible and generative. These are all considerations that guided the design of the interface presented below.

3 Learning from concordancing

Support for learner use of corpora and concordancing is premised on the fact that exposure to a word in different contexts, both lexical and grammatical, allows learners to develop a greater sense of the meaning of that word. Many features associated with using a concordancer to analyse and present word and collocation information may also lead to better retention of vocabulary items. Concordancing provides for multiple or repeated exposures, and in using a concordancer students are likely to be motivated by the 'need' to use a word – one of the three components identified as part of Hulstijn and Laufer's (2001) Involvement Load Hypothesis. Hulstijn and Laufer suggest that the involvement load is high, and therefore students are more likely to learn and retain vocabulary items if the need for particular items is determined by the learner, not the teacher. This is indeed the case if students are using a concordancer as a resource to help them improve their own writing, both to generate language items and to review ones they have already used.

Two other aspects of the Involvement Load Hypothesis, search and evaluation, are supported by concordance use. Search occurs when learners attempt to find the meaning of a word and/or how it can be used in a context. Evaluation occurs when there is "a comparing of the word with other words in order to assess whether a word … does or does not fit its context" (*op. cit.*: 14). A concordancer supports both search and evaluation.

The Involvement Load Hypothesis is not yet extensively supported empirically (see Keating, 2008). Thus we also consider the small number of studies that investigate the relationship between concordancing and learning. One of these (Fuentes, 2003) used a corpus-based approach to improve student performance in oral business English presentations. The students used two types of corpora: academic, made up of written textbook material and articles introducing basic business concepts, and professional, comprising oral business reports and product reviews. The research adopted an empirical design, with the experimental group participating in corpus-driven activities for two weeks. These activities included identifying clusters and patterns, examining a glossary, and doing fill-in-the-gap exercises. The study confirmed the positive influence of corpus-based concordancing in that the students in the experimental group produced more semi-technical business English collocations, non-business English clusters and technical compounds in their oral presentations.

Possibly of greater relevance to the present study is the work of Chambers and O'Sullivan (2004), who investigated the effect of corpus consultation in their study of eight postgraduate students writing in French. They used concordancing software particularly to improve grammatical aspects of the texts. Their teachers first underlined errors in the students' written text and placed an X to indicate basic inaccuracies such as gender, agreement, verb form, etc. Then students were asked to correct the errors by consulting the concordancer. The study showed that this helped the students identify and correct basic errors associated with gender, agreement between nouns and adjectives, capital letters in expressions such as *président de la République*, and misspelling. While most errors corrected were grammatical in nature, some could be considered to be lexical-grammatical patterning errors.

Like Chambers and O'Sullivan (2004) and a more recent study (O'Sullivan & Chambers, 2006), we sought to document the nature of errors that students could correct using the system. However, unlike them we focused on errors in collocations, and therefore present a primarily lexical perspective.

The students in the study used a concordancer that mediated a unique type of corpus of unprecedented size and scope, the Web. However as Wu, Franken and Witten (2009: 250) explain, this brings its own problems. "Web contents are heterogeneous in the extreme, uncontrolled and hence 'dirty', and exhibit features different from the written and spoken texts in other linguistic corpora" . They evaluate the capacity of the Web as corpus in terms of three features: size, cleanliness and representativeness. The fact that the Web is expansive and growing on a daily basis, that it contains language that is 'dirty' (Kilgariff & Grefenstette, 2003: 342), and the fact that is a "a highly 'skewed' archive" (Rundell, 2000: 6) all require particular attention if the Web is to be useful for learners.

4 Concordancing, the Web and the Web derived corpus

The fact that Web is a rich (but not unproblematic) source of data for linguistic analysis is evidenced by projects such as WebCorp, developed at Birmingham City University. While WebCorp initially utilized standard web search engines such as Google, its latest refinement is the design and incorporation of a search engine tailored for linguistic analysis. Renouf, Kehoe and Banerjee, developers of WebCorp,

explain its potential, as it "opens a window on text domains and types which are not available in corpora, including those which have evolved through its very existence, such as chat room talk" (Renouf, Kehoe & Banerjee, 2007: 50). They also state that "For linguists and language teachers, what WebCorp is uniquely able to provide includes neologisms and coinages; newly-vogueish terms; rare or possibly obsolete terms; rare or possibly obsolete constructions; and phrasal variability and creativity" (*op. cit.*: 50).

Several studies have explored the potential of the Web as a corpus for language learners (see, for example, Guo and Zhang, 2007; Shei, 2008). These studies have used text snippets from search engine hits to generate concordance data and discover words and word sequences in context. However, this approach is limited, because although search engines accept unlimited numbers of queries from users via Web browsers, they impose restrictions when they are accessed by other kinds of computer applications (and in some cases they prohibit such access altogether). Possibly of greater significance are features of the Web itself that make it less than suitable for language learning and teaching. These include its massive size, and the fact that it includes many items that are potentially confusing or misleading for learners, such as non-word character strings, website names and grammatical errors (see Wu, Franken & Witten (2009) for an extensive discussion of these).

Instead of relying on live Web search to generate collocation and concordance data, we work with an off-line, Web-derived corpus, the n-gram collection generated and supplied by Google in 2006. Its text was collected in January 2006 from publicly accessible English-language Web pages and amounts to approximately one trillion word tokens. The corpus contains short sequences of consecutive words, called "n-grams," along with their frequencies. The n-grams range in size from one word to five, and the 5-grams are large enough to provide useful lexical and grammatical collocation information.

Like the Web itself, this collection is large, messy, and contains anomalous language items. The n-grams must be filtered, cleaned and parsed. Unfortunately, it is virtually impossible to eliminate all grammatical errors, because of limitations in natural language processing technology. More important is the lack of context beyond the neighboring few words, which makes accurate parsing impossible in principle. Nevertheless, significant improvements can be made in this regard. We cleaned up the n-grams by using the British National Corpus (BNC) wordlist to remove non-words such as website names. We discarded word sequences if they included any words not in this list. This reduced the volume of text by 30% and yielded a much tidier corpus. The Greenstone digital library software[1] was used to organize this into a searchable digital library collection containing 145,000 words and 380 million five-grams, which was further subdivided into two collections: Web phrases, and Web collocations.

For language learners, n-grams have the intrinsic limitation that context is lost when they are removed from their original setting, a point made about corpus use in general by Widdowson (2000). Context has long been recognized as crucial for

[1] We used Greenstone version 3.03.

vocabulary learning (see Nagy (1997) for an in-depth discussion of its importance). Our remedy is to use text retrieved from two sources to reconstruct suitable contexts and present them to users on demand. This is consistent with the claim made by Charles (2007: 298) that because "students can expand the context of each concordance line at will…, [a] corpus offers the possibility of consulting the entire text and reading as much as necessary for the development of contextual knowledge".

5 Expanding the nature of the concordance data

The first source was the British National Corpus which is available from www.natcorp.ox.ac.uk. We split this into paragraphs and built them into a searchable collection, again using Greenstone. Whenever the learner asks for examples of a particular n-gram in context, we arrange for Greenstone to search the collection for occurrences and display the relevant paragraphs.

The second source is the Web itself. We wrote a program that, whenever a language learner requests the context of a particular n-gram, consults a search engine, using the words as a phrase query, and retrieves sample texts in real time. We used Yahoo as the search engine because Google disables automatic queries from computer programs other than Web browsers. Yahoo has no obvious disadvantages in terms of the quality of text snippets retrieved for a particular search.

Both sources have limitations, but the two are somewhat complementary. The British National Corpus provides far fewer examples, the number declining rapidly for four and five word sequences. In many cases there are none at all – even for items that occur reasonably frequently on the Web. For example, the string *I was very disappointed in* occurs 12,000 times in the n-gram corpus but not at all in the British National Corpus. However, when an item is found, the British National Corpus provides excellent and extensive context. Web text, being extracted from individual Web pages, is often unclean, incomplete and repetitive, as discussed above – but the examples it presents are authentic and contemporary.

6 The collections

As mentioned above, the Google n-gram corpus was organized into a searchable digital library collection and then further subdivided into two collections: Web phrases and Web collocations.

6.1 Web phrases

The phrases collection is a large subset of the original n-gram corpus. It contains about 145,000 unique words, 14 million two-grams, 420 million three-grams, 500 million four-grams and 380 million five-grams. It allows free exploration of word combinations, unconstrained by grammatical class. We build on Shei's (2008) pioneering work, mentioned earlier, which allowed users to study particular words and phrases to check whether and to what extent the text they have written represents common usage.

If users want to know what words most commonly follow a particular word or phrase they can retrieve this information. Figure 1 illustrates this for the phrase

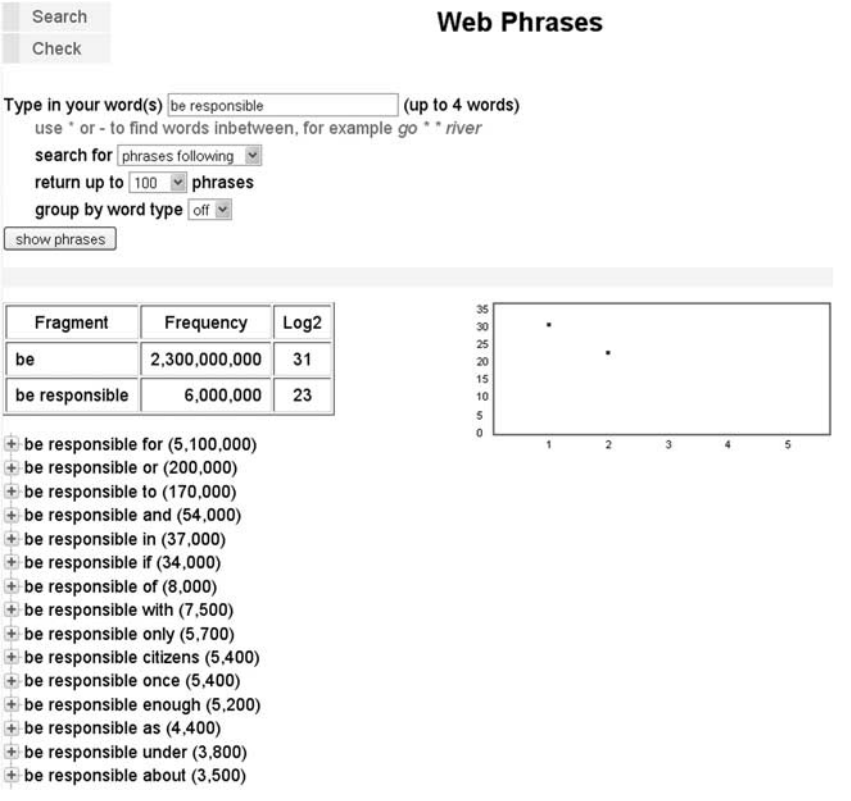

Fig. 1. Search facilities provided by the Web phrases

be responsible. The interface contains three parts. A statistical table gives the frequency count for the query term or phrase. On the right is a graph that indicates visually how the frequency (represented by its logarithm for ease of visualization) reduces as words are added. Beneath is an expandable tree that displays associated phrases in reverse frequency order, along with their frequency count.

The most frequent word following *be responsible* is *for*, then *or*, *to*, *and*, etc. Clicking *be responsible for* and *be responsible for developing*, the tree expands and displays the phrases associated with these phases, as shown in Figure 2, and the table and graph update accordingly. A phrase can be expanded up to five words, or until no further extensions are found in the collection. Samples of text that use the phrases can be retrieved from the Web and from the British National Corpus.

Users can search phrases backward by specifying the *phrases preceding* option. As shown in Figure 3, one can browse around successive words that precede *be responsible*. Most of them are modal verbs – *will*, *shall*, *would*, etc. Furthermore, an asterisk (*), which stands for any word, can be used to find words that occur between other words of a phrase. Figure 4 shows the adverbs (*solely*, *directly*, *fully*, etc.) that are associated with *be * responsible*. Further asterisks can be added, for example, *be ** responsible*, *be *** responsible,* and *be * responsible * the*.

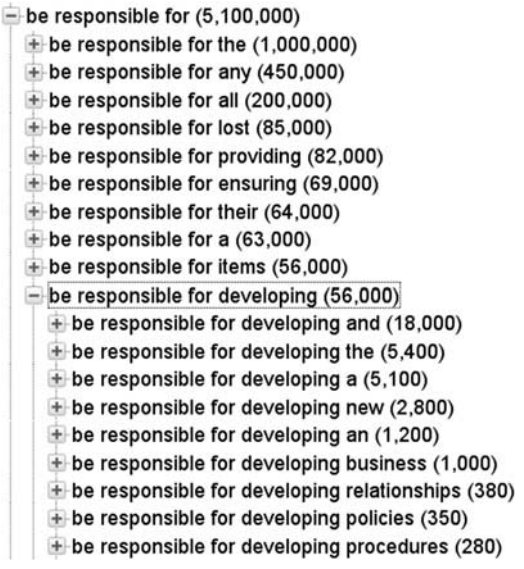

Fig. 2. Search facilities provided by the Web phrases

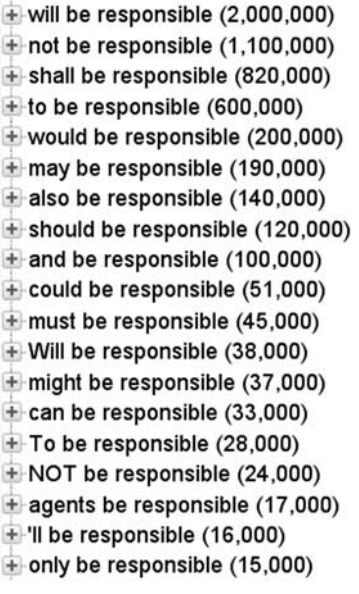

Fig. 3. Search facilities provided by the Web phrases

Finally, common words like *the, a, of,* and *to* are dominant constituents of phrases, which makes it hard for users to glean useful language patterns. To address this problem, the *group by word type* option allows users to look up the words following or preceding a phrase by part of speech of a word – preposition, verb, noun, adjective, etc.

Fig. 4. Search facilities provided by the Web phrases

6.2 Web collocations

We target ten collocation types that contain nouns, adjectives, verbs and adverbs. Table 1 gives their definition, some examples, the total number of collocations, and the number of words in this collection. We adopted six patterns (the first six) from the work of Benson, Benson and Ilson (1986) and added noun + noun, adverb + verb, verb + to + verb, and verb + adjective from the Oxford Collocation Dictionary. For nouns, determiners and possessive pronouns such as *the*, *a*, *any*, *some*, *his* are included. To make full use of n-grams, four types are extended to include further items of potential use for learners. These extensions are also shown in Table 1.

The web collocations were extracted using two- to five-grams. The identification process involved four steps:

- assign part of speech tags to the words of n-grams
- match tagged n-grams against the syntactic patterns
- discard ones that occur fewer than 100 times
- sort the extracted collocations by frequency.

The collocations are grouped by type, and search indexes are created for constituent words of a collocation. The total number of collocations and the number of words given in Table 1 indicates the sheer size of this collection. The dominant collocation types are verb + noun, noun + *of* + noun, adjective + noun, and noun + noun, each having several million examples containing around 50,000 different words. Even the smallest – verb + adjective – contains 90,000 collocations. On average, the most frequent collocations in each type occur 10 million times. If we look at a particular word, say the verb *cause*, there are 268 variants of c*ause problems*, including *cause serious problems*, *cause major problems*, *cause unpredictable problems*, etc.

Table 1 *Collocation types and examples*

collocation type	example	collocations	words
verb + noun(s) includes:	make appointments	8,700,000	54,000
verb + noun + noun	cause liver damage		
verb + adjective + noun(s)	take annual leave		
verb + preposition + noun(s)	result in the dismissal		
verb + adverb	apologize publicly	200,000	11,000
noun + noun	a clock radio	4,200,000	53,000
noun + verb includes:	the time comes	1,200,000	34,000
noun + verb with present tense	the time is running out		
noun + be + present participle	the time is spent on		
noun + be + past participle			
noun + of + noun	a bar of chocolate	7,800,000	40,000
adjective(s) + noun(s) includes:	a little girl	6,300,000	56,000
adjective + noun + noun	a solar water system		
adjective + adjective + noun(s)	a sunny beautiful day		
adjective + and/but + adjective + noun(s).	a funny and cute boy		
verb + adjective includes:	make available		
verb (incl. phrasal) + adjective	take up more	91,000	9,800
verb + noun + adjective	take it easy		
verb + to + verb	cease to amaze	440,000	11,000
adverb + verb	beautifully written	500,000	13,000
adverb + adjective	seriously addicted	200,000	10,000

To look up collocations, the user simply types in the word of interest. The system retrieves all collocation types associated with it and lets the user choose one to continue with. Figure 5 shows the result of searching for the word *cause*. First, the collocation types are grouped by word class: in this case, *cause* can be used as verb and noun. The verb section contains six collocation types related to the verb *cause*, while the noun section is dedicated to the noun *cause*. Beside each collocation type is the most frequent example of it. Clicking one, say *cause actual result*, brings up a collocation page like that shown in Figure 6. It displays more collocations of this type, sorted in inverse frequency order and presented in two columns, along with the frequency and links that retrieve samples from the BNC or the live Web. Interestingly, *cause actual result*, which occurs much less frequently in other corpora, is the top hit. This reflects the sometimes anomalous or 'dirty' nature of this Web collection, as acknowledged above (Kilgariff & Grefenstette, 2003: 342).

The user can (1) restrict the level of vocabulary displayed in the result by specifying a wordlist, (2) decrease or increase the number of collocations to return, (3) exclude collocations whose frequency falls below a particular value by adjusting a frequency cut-off, and (4) decide whether to group collocations. The first three are self-explanatory; we discuss the fourth in more detail.

Grouping is a powerful facility that allows users to inspect variants of a collocation and also helps minimize confusion caused by partial collocations. It groups collocations according to a template consisting of the main parts of a collocation

Fig. 5. Search facilities provided by Web collocation collection

type. Templates vary from one collocation type to another. For example, a verb word plus a noun word form the template of the verb + noun collocation type, while an adjective word plus a noun word is that of adjective + noun. *Cause problems*, the most common *cause* + noun collocation, has 268 variations. *Cause serious problems*, *cause unpredictable problems*, *cause major problems*, etc are grouped under the cause + problems template. Ones for cause + side include *cause side effects, cause different side effects, cause exaggerated side effects*, etc.

The user enters two words and configures the parameters described above. The system retrieves the collocations associated with these words, groups common and

cause actual results	760,000		cause problems	750,000	
cause damage	300,000		cause cancer	280,000	
cause death	260,000		cause injury	240,000	
cause harm	210,000		cause pain	180,000	
cause confusion	170,000		cause a problem	140,000	
cause drowsiness	140,000		cause side effects	130,000	
cause dizziness	120,000		cause trouble	120,000	
cause a lot of	120,000		cause a denial of	110,000	
cause different side effects	110,000		cause a denial	110,000	
cause disease	110,000		cause irritation	110,000	
cause serious problems	110,000		cause birth	98,000	
cause any problems	96,000		cause harmful interference	95,000	
cause stomach	92,000		cause serious injury	89,000	
cause a lot	88,000		cause a delay	88,000	
cause birth defects	87,000		cause an increase	80,000	

Fig. 6. Search facilities provided by Web collocation collection

Common collocates between speak and tell:

speak the truth	230,000		tell the truth	920,000	
speak for anyone	63,000		tell anyone	820,000	
speak unto thee	17,000		tell thee	110,000	
speak with people	13,000		tell people	1,300,000	
speak for the rest	13,000		tell the rest	33,000	
speak with your doctor	11,000		tell your doctor	360,000	
speak of things	11,000		tell things	26,000	
speak for the whole	11,000		tell the whole	170,000	
speak a lot	10,000		tell a lot	63,000	

Different collocates of speak and tell:

speak on behalf of	300,000		tell millions of	2,400,000	
speak the language	230,000		tell a friend	850,000	
speak on behalf	180,000		tell the difference	770,000	
speak a language	170,000		tell the story	660,000	
speak about curriculum	120,000		tell the story of	580,000	
speak the same language	100,000		tell the world	430,000	
speak a word	98,000		tell your friends	420,000	
speak the language of	85,000		tell a story	410,000	

Fig. 7. Search facilities provided by Web collocation collection

different ones together, and presents them side by side. Figure 7 shows the result of comparing the verb *speak* and *tell* in the verb + noun type. These have 11 out of 100 collocations in common. The most frequent collocations are *speak on behalf of* and *tell millions of* respectively. *Speak* and *tell* can both be used with *truth*, *someone*, *everyone*, *anyone* etc. However, with only two exceptions, *tell* collocations are far more frequent than *speak* ones. The latter verb tends to be associated with a preposition when the noun is a person. If we look at the different collocations, *speak* and *tell* have quite different usages. We say *speak a language*, *speak my mind*, *speak a word*, *speak ill* (or *evil*) *of someone*, whereas we say *tell the difference*, *tell a story*, *tell the time*, and (mostly) *tell someone*.

6.3 Designing a users' guide

We designed a users' guide for the two collections based on samples of student text that are included as exemplars in the IELTS Specimen Materials Handbook (IELTS, 1997). By analyzing typical errors that students make, and relating them to the possibilities that the system affords, we created five kinds of exercise. Here we give a brief description of the guide.[2]

First, the system is useful for essay preparation. Given a topic, say *nuclear power*, students can use the system to find appropriate vocabulary in two ways. They can collect useful noun + noun, adjective + noun or noun + *of* + noun phrases using topic-related keywords like *nuclear*, *weapons*, *energy*, *benefits*, *threat*, *disadvantages*, *solutions*. They can also learn what verbs are commonly associated with those words, and their correct usage. For example, we say *pose a threat*, not *give threat*; *the benefits outweigh the disadvantages*, not *we outweigh the benefits and disadvantage*; *find solutions*, not *examine about the solutions*.

Second, learners tend to reuse particular words repeatedly throughout their essays, because of limited vocabulary knowledge. A typical example is overuse of the verb *rise* or *decline* in the IELTS task that asks for a description of changes and trends in an input text, graph, table or diagram. Examining collocations of words like *shares* or *prices* will quickly yield alternatives such as *jump*, *soar* and *surge*; or *drop*, *fall*, *slump*, *slip* and *plunge*.

Third, learners often misunderstand the usage of a word, and overgeneralize common words like *have*, *do*, *make*, *take,* and *give*. As a result, odd word combinations or idiosyncratic word choices are scattered throughout their writing. Examples are: *cultivate their children with*, *reinforce the income*, *deep interests, give threat*, *the city must have another solutions*. The collocation collection can help learners make more accurate or appropriate choices of words and word sequences. For example, we could ask students to look up the nouns that follow *cultivate*, or find verbs that are commonly associated with *solutions*.

Fourth, learners also find it difficult to boost or hedge statements by adding adverbs. Suppose one wants to add extra strength to the sentence *We will all benefit from it*. Searching *benefit * from* in the Web phrases yields *greatly*, *directly*, *significantly*, *enormously* and *immensely*. Or consider how adverbs are used to describe feelings appropriately and precisely. If one wishes to express disappointment, the Web phrase collection provides a wide range of modifiers, from *extremely*, *deeply*, *bitterly*, *pretty*, *quite* to *rather*, *somewhat*, *just*, *slightly*.

Finally, we designed exercises to demonstrate how to use the system to correct grammatical errors. Misused prepositions and ill-formed verbs were two dominant grammatical errors in the sample text: for example, *The government must be responsible <u>of</u> their welfare*, *They have increased day <u>to</u> day* and *this problem would <u>resolve</u> a little*. Those errors can be corrected by searching the Web phrases for *must be responsible, increased day * day* and *this problem would*.

[2] The full version is available at http://flax.nzdl.org/greenstone3/instruction.htm.

7 Evaluation

We conducted two types of evaluation to assess the utility and effectiveness of the two collections, and the way in which they can be used to generate useful language examples to improve text. First, one of the authors (an expert user) used the system to attempt to resolve errors in students' writing to discover its potential to offer correct, appropriate and accessible alternatives. Then we asked language students to use it in conjunction with the guide, so that we could evaluate the use they made of it and how it affected their textual revisions.

7.1 The students

The researchers worked with teachers in our institution's language support centre to recruit participants. The researchers targeted students who were involved in the IELTS writing preparation class. Nine language learners, three females and six males, from 18 to 30 years old, and native speakers of five different languages, volunteered to participate in the evaluation.

7.2 Generating and preparing student texts for evaluation

During the first session, the students were given an IELTS argument writing task[3] selected by their teacher as part of their normal class program. They were asked to write a response to the task within the usual forty minutes time allocation. However, contrary to normal practice, they were asked not to use dictionaries.

After this, one of the researchers who is very familiar with the system and an experienced teacher examined the students' writing, highlighting aspects of the texts they felt needed improvement and revision. It should be noted that while we have labelled these as 'errors', in many cases they are examples of not quite acceptable words or word sequences. While this seems a broad brief, as guidance, two areas were suggested for focus:

1. grammatical errors, e.g., incorrect use of verb forms and prepositions, misused plurals and articles, and missing verbs.
2. lexical errors, e.g., wrong or inappropriate word combinations, particularly those involving noun + verb, verb + noun, adjective + noun and noun + noun combinations.

Consistent with the approach taken by Chambers and O'Sullivan (2004), the text was highlighted at the phrase level. For example, in the student's text below, the brackets [ ] indicate the phrases identified as needing to be revised.

Some famous museums have become [one the most powerful attractions] to [reinforce the income] for a particular country.

The teacher and researcher met to compare marked sections of text. When agreement was reached, additional marking to help students focus on particular parts of the highlighted phrases was added, if appropriate. For example, in the

[3] The task was: Historical art has more cultural value than modern art. Discuss both sides of this argument and give your opinion.

following text, the words *powerful* and *reinforce* were underlined to assist student searching of collocations, and the symbol ^ was used to indicate a missing element.

> Some famous museums have become [one ^ the most] [powerful attractions] to [reinforce the income] for a particular country.

7.3 Assessing the system's potential

As explained above, before evaluating the students' use of the system to make changes to their text, we used the system to check the errors ourselves with the aim of establishing baseline data. The evaluation was conducted by one author who is very familiar with the system and is also a second language learner.

The errors in students' texts were identified in six types of structures: noun phrase, verb + noun, noun + verb, prepositional phrase, phrasal verb or verb + preposition, and verb + complement. Another large group of errors were classified as grammatical as they involved morpheme omission or error. Table 2 shows the frequency of these errors, both within these categories and within the subcategories associated with some of the categories. It also gives examples of acceptable alternatives generated by the system.

In total, 108 errors of all types were identified across students' texts. The system was able to generate correct and appropriate alternatives in 95 (just under 88%) of cases. If we focus specifically on lexical non-grammatical errors, the success rate is higher, with 82 corrections (just under 94%).

Errors associated with the noun phrase (adj. + noun, and noun *of* noun), together with errors in the verb + noun pattern, were the most frequent. If we combine sequences involving preposition use, that is, preposition phrases, phrasal verbs and verb + preposition, the result is 17 errors, a smaller but still substantial number. Of these 17, only two errors were not resolved.

Grammatical errors represent a large group (27), but in contrast to the success of the system in resolving more lexical errors, a relatively large number of the grammatical errors were not resolved.

7.4 Student use

Having marked up their text as described above, we gave the students the users' guide. The system was explained to them in more detail and they were taken through the guide with demonstrations. This was undertaken in a two-hour session.

In a second two-hour session students were asked to revise their text using the system, focusing particularly on the marked-up sequences. Their actions were logged automatically for later analysis. A third session was available for students who needed more time to complete their revisions.

Table 3 indicates how the system was used. It shows the number of sequences that were marked up; the number of those that students attempted to change; and what percentage of those changes were successful and unsuccessful. For instance, in the case of errors associated with the noun phrases of the adj. + noun form, 18 sequences were marked up, of which the students changed 13 (72%) successfully and 5 unsuccessfully. The high success rate indicates the students' willingness and ability to use the system to revise their work.

Table 2 *System generated alternatives to students' errors with collocation categories*

Total frequency of system generated alternatives		Frequency within each category		Frequency within each subcategory		Examples	
						student text	system generated alternatives
Noun phrase	36 (2)	adj. + noun	19 (1)	adj. *	3	contemporary arts building	contemporary art gallery
				* noun	16	a fancy and good position	a unique position
		Noun *of* noun	17 (1)	* of noun	14	the most important steps of our evolution	stages of evolution
				noun of *	2	a element of a national spirit	an expression of national spirit
				noun ** noun	1	important events in their times	events of that time
Verb + noun	27 (2)			* noun	26	reinforce the income	increase the income
				verb *	1	help common people	help ordinary people
Noun + verb	3			* verb	1	the essay favour	I favour
				noun *	2	the profound influence created by	the profound influence exerted by
Prep. phrase	8 (1)					in the other hand	on the other hand
Phrasal verb; verb + prep.	7 (1)					play an important role on	play an important role in
Verb + complement	4					the argument may be true	the argument may be valid
Adverb use	3					are aware of a lot	are fully aware of
Grammatical errors	20 (7)					more likely to be preserve	more likely to be preserved
	108 (13)						

Note: Brackets indicate additional errors that were unresolved in our analysis.

Table 3 *Student changes to errors identified in their texts within collocation categories*

		Number of marked word sequences changed	Frequency of changes within each category	Successful changes (%)
Noun phrase	Adj + noun	18	13 (5)	72
	Noun *of* noun	18	14 (4)	78
Verb + noun		27	16 (11)	59
Noun + verb		3	2 (1)	67
Prep phrase		8	5 (3)	62
Phrasal verb; verb + prep		7	7	100
Verb + complement		4	3 (1)	75
Adverb use		3	2 (1)	67
Grammatical errors		20	11 (9)	55
Total		108	73 (35)	67

Note: Brackets indicate changes that led to anomalous or grammatically incorrect text.

Adj. + noun and "noun *of* noun" both showed a consistent and relatively high success rate. In most cases students used correct main nouns (the second component of noun + noun), but picked inappropriate adjectives and modifying nouns, resulting in strange combinations – for example, *main culture value*, *powerful attractions*, *classical artifacts*, *numerous of countries*, *a great deal of museum*, *these sort of arts*, and *popularity of modern technology*. Students obtained good results on this kind of error, but the success rate declines when both parts are wrong. As an encouraging example in the "noun *of* noun" category, one student changed *modern art's appearing* to *the development of modern art*.

Students performed particularly well (100%) on the verb + preposition category, owing, we believe, to many useful examples the system provides. For instance, they changed *play an important role on* to *play an important role in, give priority for* to *give priority to* and *is famous with* to *is famous for*.

Verb + noun is challenging, as changing the verb may alter the meaning of the whole sentence. The system can give the verbs that are most frequently associated with a particular noun, but it is up to the student to pick an appropriate one. Some students chose ones that they were most familiar with regardless of context, which was not necessarily the best choice. Sometimes they chose one that made a good verb + noun combination but did not fit the context.

The result of successful changes in the grammatical errors category is largely consistent with the success rate in other categories, though slightly lower at 55%. For the other categories, the data size is too small to give us much of a sense of the pattern of changes. However, examples of changes made in some of these categories look to be very successful. For example in the category verb + complement, *society has become more increasing fascinating* was changed to *society has become more accepting*; and *has made the society become more valuable* was changed to *has made the society become more open and liberal*. In the adverb changes, the following

example, a change from *modern people strongly claim that* to *modern people legitimately claim that*, indicates the potential of system use to provide students with examples that appear very 'native-speaker-like'.

Of the 108 marked sequences, we can only identify 95 changes, either successful or unsuccessful, in the students' text. Thirteen marked sequences were not used in the revised version – in other words they were abandoned. This represents a type of avoidance strategy. This happened in particular with one student who discarded the seven sequences and rewrote a substantially different text from her draft. The log data showed that the students actually did some work on all 13 sequences, but gave up after a few unsuccessful attempts. We treat those removed sequences as unsuccessful changes, although sometimes removing them improved the text. In total, the student success rate was 67% (73/108) – 70.5% if grammatical errors are excluded. Compared with our assessment of what is possible using the system, the students achieved a 77% (73/95) success rate on their own. This gives a strong indication of the willingness and ability of the students to use it for text revision.

7.5 Discussion

One of the major limitations of the study is the time allocated to the evaluation. An in-depth study to capture students' perceptions and strategies while using the system is clearly needed. Nonetheless, within the present study we can make the following observations. When use of the system resulted in a modification to the text, the alteration was most often an improvement, although some local changes did not necessarily produce better text overall. However, the system certainly has potential for helping students make correct and more appropriate word choices, and thereby generate more correct, appropriate and 'native-like' word sequences or collocations. The frequency-based phrases that it provides help students focus on actual usage of particular words, including nuances that are generally left unarticulated in language teaching.

One aspect that we have not touched is to use the system to help generate the text, which is addressed in the first part of the user guide. However, as we worked through the students' writing we noticed the low volume of noun phrases. In particular, occurrences of *noun of noun* were limited to quantification words such as *number*, *a great deal* and *lot*. In fact, this particular phrase type is prominent in academic writing, and we believe the system will help students improve their collocation knowledge in this respect.

There are several limitations of the system that need to be addressed; some are attributable to the learners themselves and their level of proficiency. Students have to pick a correct word or phrase to obtain a reasonable result, which is often difficult for learners. The choice of word form (such as singular or plural) and the presence or absence of an article may yield substantially different results. Another issue is that function words like *the, a, of,* and *to* are dominant constituents of phrases, which makes it hard for users to glean useful language patterns.

With respect to the system itself, the learners may be overwhelmed by the massive amount of data the system provides. It may be advisable to build sub-collections for particular user groups. In addition, these collections are based on a historical dump

of the Web, and have been further filtered: as noted earlier this falsely rejects some acceptable phrases – such as ones containing neologisms like *google*. Finally, grammatical errors in Web text may confuse less advanced learners, and the situation is aggravated when they occur reasonably frequently. For example, *may not suitable* occurs 602 times in the collection. Taking all this into account, it is important to provide training on the use of the system.

8 Conclusion

The area of word combinations is particularly important to learners. As Nesselhauf (2003: 223) explains, ''Collocations are of particular importance for learners striving for a high degree of competence in the second language but they are also of importance for learners with less ambitious aspirations, as they not only enhance accuracy but also fluency''.

The system we have described and evaluated makes use of digital library software and its search and retrieve functions to present information to learners about word sequences, or collocations. While the system still needs refinement, the information it presents is arguably better than that offered by many concordancers in that it is frequency based and therefore prioritizes certain patterns over others. Although it presents items in a limited context, learners have the capacity themselves to browse more extended contexts. It is not limited to a restricted set of syntactic patterns, but encompasses a wide range (based on Benson, Benson, & Ilsen, 1986), and the collection is searchable by collocation type.

In order to provide a realistic context of use, we recruited nine language learners from an IELTS writing preparation class. Each wrote an essay, in which we examined each language error and determined whether, in principle, the system could help resolve it. Then we marked the position of the errors and asked the students to use the system to correct them. Results were extremely encouraging. Of a total of 108 errors, the system could in principle help resolve 95 of them and the students actually resolved 73 without any human assistance.

We recognize the limitations of an evaluation that is conducted for just a short period of time. However, we believe that an automated system that measurably helps learners use word combinations more accurately and appropriately, even within these constraints, is surely worthy of further development and evaluation.

References

Benson, M., Benson, E. and Ilsen, R. F. (1986) *The BBI combinatory dictionary of English: A guide to word combinations*. Amsterdam/Philadelphia: John Benjamins.

Bishop, H. (2004) The effect of typographic salience on the look up and comprehension of unknown formulaic sequences. In: Schmidt, N. (ed.), *Formulaic sequences: Acquisition, processing, and use*. Philadelphia, PA, USA: John Benjamins Publishing Company, 227–244.

Chambers, A. and O'Sullivan, Í. (2004) Corpus consultation and advanced learners' writing skills in French. *ReCALL*, **16**(1): 158–172.

Charles, M. (2007) Reconciling top-down and bottom-up approaches to graduate writing: Using a corpus to teach rhetorical functions. *Journal of English for Academic Purposes*, **6**(4): 289–302.

Cobb, T. (n.d.) *Compleat Lexical Tutor*. http://www.lextutor.ca/

Cobuild Concordance and Collocations Sampler. http://www.collins.co.uk/Corpus/CorpusSearch.aspx

Fuentes, C. A. (2003) The use of corpora and IT in a comparative evaluation approach to oral business English. *ReCALL*, **15**(2): 189–201.

Gabrielatos, C. (2005) Corpora and language teaching: Just a fling or wedding bells? Teaching English as a second or foreign language, 8(4), http://tesl-ej.org.ezproxy.waikato.ac.nz/ej32/a1.html

Greenstone Digital Library Software. http://www.greenstone.org

Guo, S. and Zhang, G. (2007) Building a customized Google-based collocation collection to enhance language learning. *British Journal of Educational Technology*, **38**(4): 747–750.

Hulstijn, J. H. and Laufer, B. (2001) Some empirical evidence for the involvement load hypothesis in vocabulary learning. *Language Learning*, **51**: 539–558.

International English Language Testing System (IELTS) (1997) *Specimen materials handbook*. http://www.scribd.com/doc/13570277/

Keating, G. D. (2008) Task effectiveness and word learning in a second language: The involvement hypothesis on trial. *Language Teaching Research*, **12**(3): 365–386.

Kilgariff, A. and Grefenstette, G. (2003) Introduction to the social issue on the web as corpus. *Computational Linguistics*, **29**(3): 333–347.

Nagy, W. E. (1997) On the role of context in first- and second-language vocabulary learning. In: Schmitt, N. and McCarthy, M. (eds.), *Vocabulary description, acquisition and pedagogy*. Cambridge: Cambridge University Press, 64–83.

Nation, P. (2001) *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nesselhauf, N. (2003) The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, **24**(2): 223–242.

O'Sullivan, Í. and Chambers, A. (2006) Learners' writing skills in French: Corpus consultation and learner evaluation. *Journal of Second Language Writing*, **15**: 49–68.

Peachey, N . (2005) Concordancers in ELT. In: British Council teaching English. http://www.teachingenglish.org.uk/think/articles/concordancers-elt

Renouf, A., Kehoe, A. and Banergee, W. (2007) WebCorp: An integrated system for web text search. In: Nesselhauf, C., Hundt, M. and Biewer, C. (eds.), *Corpus linguistics and the web*. Amsterdam: Rodopi, 47–68.

Rundell, M. (2000) The biggest corpus of all. *Humanising Language Teaching*, **2**(3): http://www.hltmag.co.uk/may00/idea.htm

Shei, C. C. (2008) Discovering the hidden treasure on the Internet: using Google to uncover the veil of phraseology. *Computer Assisted Language Learning*, **21**(1): 67–85.

Stubbs, M. and Barth, I. (2003) Using recurrent phrases as text-type discriminators: A quantitative method and some findings. *Functions of Language*, **10**(1): 61–104.

Webcorp. http://www.webcorp.org.uk/index.html

Wei, Y. (1999) *Teaching collocations for productive vocabulary development*. (Report No. FL 026913). Developmental Skills Department, Borough of Manhattan Community College, City University of New York. (ERIC Document Reproduction Service No. ED457690).

Widdowson, H. G. (2000) On the limitations of linguistics applied. *Applied Linguistics*, **21**(1): 3–25.

Wu, S., Franken, M. and Witten, I. H. (2009) Refining the use of the web (and web search) as a language teaching and learning resource. *Computer Assisted Language Learning*, **22**(3): 249–268.